DOCUMENT RESUME

ED 021 324

EA 001 461

By- Aikin, Marvin C.; Duff, William L., Jr.

SOME DATA PROBLEMS IN SYSTEMS RESEARCH.

Note- 11p.

EDRS Price MF-$0.25 HC-$0.52

Descriptors- *DATA, *DATA COLLECTION, DATA PROCESSING, *DECISION MAKING, MODELS, *PROBLEM SOLVING, *SYSTEMS ANALYSIS

The process of dealing with real wor'd problems often requires solutions that would not be totally acceptable in the world of "pure" scientific research. A systems analytic viewpoint demands that everything possible be done to define the constraints of the system and that the procedures used to construct a model out of imperfect data be specified. Four types of problems are encountered in doing systems research: (1) specification of output measures, (2) data incompatibility where different instruments are used, (3) data incompatibility where the concept is not precisely captured by existing data, and (4) missing data. Procedures for solving the problems in each of these problem areas are suggested. (HW)

SOME DATA PROBLEMS IN SYSTEMS RESEARCH

by

Marvin C. Alkin
Assistant Professor of Education
University of California, Los Angeles

and

William L. Duff, Jr.
Research Associate, Institute for Development of Educational Activities, Research and Development Division, Los Angeles.

> "When I use a word," Humpty Dumpty said in a rather scornful tone, "it means just what I choose it to mean--neither more nor less."
>
> "The question is," said Alice, "whether you can make words mean different things."
>
> "The question is," said Humpty Dumpty, "which is to be the master--that's all."
>
> Through the Looking Glass
> Louis Carroll

It would certainly be convenient if the researcher could transform the meaning of data with the impunity Humpty Dumpty enjoys in transforming words. If this were permitted, some of his more troublesome problems would mercifully vanish through Alice's looking glass. But, alas, the researcher is disciplined by the requirements of his trade. He is expected to reserve certain words for certain uses, and when he uses a word to describe or identify different concepts or things, he is obligated to explain and justify his actions.

"Systems" is a Humpty Dumpty type word; its meaning depends a great deal upon who uses it. Some speak of philosophical systems while others are concerned with control systems, and still others with political systems or weapons systems. Some use systems to identify a physical construct w ile others use it to describe a conceptual approach. Many use the term to indicate their concern for complicated organizations made up of many interrelated parts. Each has equal claim to the word. The writers, however, define a system in rather simple terms: A system is an entity that has at least one input and at least one output associated with it; it may or may not be complicated. Furthermore, we restrict our interest to only those systems that are controllable by man. In the remainder of this paper the use of the term "system" will be consistent with this definition.

The systems analyst is usually concerned with building a model of a "real world" system. A model is an abstraction from, and a simplification of, reality, which hopefully, captures the crucial relationships in the real world. Systems are often enveloped by larger systems. For example, a classroom, a school, a district, a state's or nation's educational facilities can be legitimately defined as systems. The delineation of a specific system depends upon the decisions one wishes to make and the related questions the analyst wishes to answer. The generic question the systems analyst attempts to answer is, how can we maximize the systems output utilizing available resources. To this end he must evaluate the resource cost and the corresponding outputs associated with various combinations of inputs.

During the past years we have attempted to apply systems research to educational problems. Basically, our concerns have been with the decision options of administrators. Thus, a question of concern is, how can school principals and superintendents modify the manner in which they use resources within their institutions in order to maximize educational outputs. To accomplish these purposes, we have initiated a number of systems studies, one of which involve high schools in California.

We developed a mathematical model simulating a high school in terms of a description of its inputs (student and school), selected outcomes, and various organizational characteristics believed to mediate in the achievement of the outcomes. As a first step, multiple regression techniques were used to identify the relationships between the inputs and outputs. The inputs were divided into two categories--those which are controllable by school administrators and those which are not. While "uncontrollable" inputs are important in that they interact with and thereby affect any decisions relating to the "controllable" inputs, our main interest is in the examination of the administratively controllable variables.

Included among the "uncontrollable" inputs were a variety of data items descriptive of the socioeconomic and student characteristics of the school environment. "Controllable" inputs included items descriptive of teacher characteristics, school programs, and organizational characteristics of both the school and the district.

In this paper, we propose to discuss four types of data problems that we have encountered in doing systems research. The first of these is related to the difficulties of specifying outputs. The specification of the output measures is perhaps the most difficult problem in systems studies. It is obvious that there is nothing approaching a consensus in defining the specific objectives of educational systems. They certainly are not given in a usable way by a national, regional, or even local or district authority. In addition, after reviewing the rather nebulous statements of objectives that did exist, it became apparent that they were often multiple and conflicting. Under these conditions it was obvious that alternative means (inputs) used to reach any one end (output) of the system would cause "spillover" effects—both negative and positive—on other ends of the system. It should be mentioned that this is not necessarily a criticism of school administrators, district boards or others. Indeed, it is an almost unavoidable concommitant in studies of complicated systems, such as public education.

It is impossible to define meaningful objectives for systems studies without knowing something about the feasibility and cost of reaching them. We consider, therefore, that the formulation of and learning about objectives, is a prime purpose of systems studies. For all of these reasons, systems studies are an iterative process; assumptions necessary to specify the model as well as the criterion measures must be derived from, and played back against, the analysis. This is best illustrated by relating a story told by Charles Hitch in one of his early papers on systems analysis:

> A friend of mine who is a sophisticated systems analyst once tried to solve a personal problem by a rigorous maximization of an objectives function supplied by his doctor. He needed to lose weight, so he determined by consulting the experts his minimum requirements for proteins, carbohydrates, fats, vitamins, minerals, etc. He also obtained the quantities of each of these food elements in the 500 or 600 foods on the BLS list. Then, on the plausible theory that mass is filling and that most dieting attempts fail because the subject feels hungry, he maximized, subject to various constraints, the weight (not counting water content) of the diet that would give him his minimum caloric requirements. The answer, ignoring minor quantities of various foods, was that he should drink 80 gallons of vinegar per day (vinegar is a weak acid, and its weight per calorie is remarkably high). Since his own taste buds and digestive tract were to be the victims of this experiment, he knew intuitively that the answer was crazy and informed his machine that it should recalculate, ignoring vinegar. The second answer, incidentally proved to be as unacceptable as the first, so he introduced still other conditions.1/

Now, Hitch's colleague was proceeding very sensibly with his problem, and properly using the tools of the systems analyst. But a part of the process was being able to recognize what is a reasonable solution, and having the ability to introduce complications and constraints as their necessity became apparent. Hitch underscores this observation later in the same paper.

> .....it is slightly worrisome that the method used is very similar to the one so many of us use to take some plausible objective as given, and calculate like mad to maximize it. But we are using it in areas where our intuition doesn't reach very powerfully, and it therefore isn't so easy to recognize vinegary answers for what they are. That doesn't keep them from being just that.2/

It is obvious, from Hitch's story, that learning about outputs is one of the chief outputs of systems studies. We are reminded to look at our outputs as carefully as we look at our

1. Hitch, Charles, J., On the Choice of Objectives in Systems Studies, The RAND Corporation, Economics Division, March 30, 1960, p. 9.

2. Ibid., p. 10.

model and its inputs. If we begin with tentative objectives, such as we did in our high school study, we should expect to replace or modify them as we move along. It is unlikely that we will be able to define satisfactory objectives at the beginning of a study.

The purposes of a system study, therefore, are twofold: a) to provide information for rational administrative decision making in order to improve the real world system, and b) to continually re-examine and modify the arguments of the model in order to improve the model and the analysis.

Another problem in systems research that we have faced in our study is that of data incompatibility. Unlike the researcher in the laboratory, the systems analyst must often use data collected by other people for other than its original purposes. Frequently, this data is not wholly suited to his ends. He is nevertheless often forced to use it due to the expense involved in gathering the large quantity of information usually needed in systems studies. In specifying our high school model, for example, we included over 200 variables from each of 180 high schools. Thus, the study, had it been based on data collected especially for this purpose, and had it included responses from each of the students involved, would have required in the neighborhood of 80 million individual observations. In addition to the direct cost of collecting the data we must add the indirect costs that we might, under other circumstances, shift to the school and its students. The loss of student and faculty time is no less a cost in data gathering than is the cost of printing the testing instruments. Faced with the total costs of recollecting compatible data, we searched for reasonable alternatives.

Conceptually, we identified two major categories of data incompatibility. We may think of these as the second and third problems to be discussed in this paper. The first occurs when different instruments are used to measure the same concept. The second occurs when the concept we wish to measure is not precisely captured by existing data categories. In our study we have used or considered several different methods of coping with these problems.

The problem in the first type of data incompatibility is to find an equation that accurately transforms one information piece to another. For example, if we found that some schools recorded the body weights of students in grams and others recorded them in ounces, we could simply multiply the gram weight by 28.35. By doing so, all the data are transformed to ounces and are compatible. Thus, if we found that some of our subject population had taken different mathematics achievement tests, we might (after taking a deep breath and crossing our fingers) be willing to postulate a relationship between the recorded scores of the two groups. A second alternative is to simply throw away the suspect data. In our study, we often chose this latter alternative. (As a matter of fact, we discarded over 33 percent of the sample for this or related reasons.)

In cases where we felt relatively sure of the conversion formula, we attempted the transformation. For example, in our study we found that each high school listed the scores of entering students (8th grade) for the areas of reading and arithmetic and also 11th grade scores for the same areas. Three of the summary scores reported were: median score, score at the 1st quartile and score at the 3rd quartile. For comparison purposes, scores on different tests had to be converted to one comparable form. Before we could proceed with our study, the above three summary scores were converted to percentile scores on national norms for each of the tests. Thus we, in effect, had produced three new data items which were the national percentile scores for the student at the median, first, and third quartiles at each school.

The identification of the second type of data incompatibility is a rather subjective exercise. A great deal of what we call real data is in fact a proxy for some abstract concept. Consider, for example, the problem we faced in measuring one of our criterion concepts. We wished to identify the effects that the inputs have on college attendance and performance. The input data, however, was on students still in high school and it was impossible to get output measures on these same students without waiting several years. Our solution was to use as an output measure the attendance and performance of preceding graduating classes from each school, the assumptions being that: a) the nature of the community and, consequently, the student input to the system (i.e., the uncontrollable variables) have remained relatively constant, and b) the educational program (i.e., the controllable variables to which the graduates were subjected) are substantially the same as that which presently exist. These two assumptions seem reasonable in view of the fact that schools and communities generally are slow to change.

In addition, we also found that some already available data would be of more use if it were combined or changed in certain ways. We felt that an examination of students attending college and their success is more properly expressed as a function of academically able students in the school population. For example, by taking the ratio of bright students (those with an IQ of 115 or above) to those attending college, we obtained a more suitable criterion measure.

Up to this point we have discussed rather briefly, the criterion problem and the two types of data compatibility problems. We shall now turn to a fourth problem area--missing data. We recognized that schools varied in the consistency with which they record information. The schools simply did not record all scores on all students. This was really no surprise, but it nevertheless presented us with some messy problems. As we have noted, in cases where the data was missing from an output measure, we decided to discard that case from the analysis. With regard to input categories, on the other hand, we attempted to statistically recapture the missing information.

The procedure we used was really quite simple. In statistical terms, we were faced with the situation of having a different number of observations represented in specific zero-order correlations between predictors and criterion variables. The problem was solved, in effect, by uniting two existing computer programs available at the Health Sciences Computing Facility at U.C.L.A. The two programs are: a) the BMDO3D program which produces an intercorrelation matrix with missing data (i.e., individual zero-order correlations are based only on those observations for which data is present); and b) the BMDO2R program which is a stepwise linear multiple regression equation. Thus, we developed a program which uses the intercorrelation matrix produced by BMDO3D as the input for a stepwise regression equation. We consider this technique more appropriate for our purposes than the traditional practice of filling missing data items on individual observations with the mean of that item—a procedure which tends to reduce the variance on the item.

In addition, we are presently engaged in examining the possibility of developing prediction models for each of the independent variables in terms of the other predictors, and using resultant equations to generate estimates of missing data on an individual case basis.

We have mentioned but four of the many problems facing the systems researcher. This was partly because these are, in our judgment, some of the more interesting and pervasive problems that we have encountered. It should be remembered, however, that we are in the first stages of our study and, therefore, suspect the existence of many problems that we haven't as yet identified, much less solved.

We are proceeding with the abandon of one who has a problem to solve but knows he may never have all the information necessary to find a definitive solution. We lament the necessity of subjecting our data to what some might call "Humpty Dumpty transformations," but we harbor no delusions of impunity. The process of dealing with real world problems, often-times requires solutions that would not be totally acceptable in the world of "pure" scientific research. A systems analytic viewpoint demands thaւ we do what is possible to define the constraints of the system and that we specify the procedures we have used to enable ourselves to construct a model out of imperfect data. We have defined four kinds of problems: (1) specification of output measures, (2) data incompatibility where different instruments are used, (3) data incompatibility where the concept is not precisely captured by existing data, and (4) missing data. For each of the problem areas we have indicated the procedures we used to solve the problem. We recognize the imperfections in the proposed problem solutions but hope to resolve some of these difficulties by "testing for vinegar" at each stage of the analysis.